COMP90042

Web search and text analysis

Workshop Week 7

Your tutor

- Winn Chow (Senior Tutor)
- winn.chow1@unimelb.edu.au
- Office: Doug McDonell - 9.23
- Here, you can find my workshop slides:
- https://github.com/winnchow/COMP90042-Workshops

Q1

1. What is a **POS tag**?

   (a) What are some common approaches to POS tagging? What aspects of the data might allow us to predict POS tags systematically?

   (b) POS tag (by hand) the following sentence: `Pierre Vinken, 61 years old, will join the board as a nonexecutive director Nov. 29.` according to the Penn Treebank tags. (Note that some of the tags are somewhat obscure.)

POS Tagging

Tagged text Example

The/DT limits/NNS to/TO legal/JJ absurdity/NN
stretched/VBD another/DT notch/NN this/DT week/NN
when/WRB the/DT Supreme/NNP Court/NNP
refused/VBD to/TO hear/VB an/DT appeal/VB from/IN
a/DT case/NN that/WDT says/VBZ corporate/JJ
defendants/NNS must/MD pay/VB damages/NNS
even/RB after/IN proving/VBG that/IN they/PRP
could/MD not/RB possibly/RB have/VB
caused/VBN the/DT harm/NN ./.

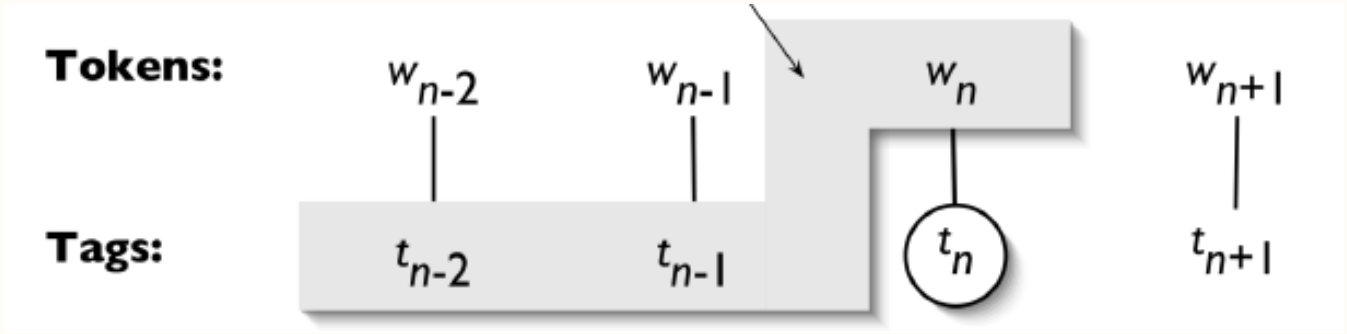
Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	*and, but, or*	PDT	predeterminer	*all, both*	VBP	verb non-3sg present	*eat*
CD	cardinal number	*one, two*	POS	possessive ending	*'s*	VBZ	verb 3sg pres	*eats*
DT	determiner	*a, the*	PRP	personal pronoun	*I, you, he*	WDT	wh-determ.	*which, that*
EX	existential 'there'	*there*	PRP$	possess. pronoun	*your, one's*	WP	wh-pronoun	*what, who*
FW	foreign word	*mea culpa*	RB	adverb	*quickly*	WP$	wh-possess.	*whose*
IN	preposition/ subordin-conj	*of, in, by*	RBR	comparative adverb	*faster*	WRB	wh-adverb	*how, where*
JJ	adjective	*yellow*	RBS	superlatv. adverb	*fastest*	$	dollar sign	$
JJR	comparative adj	*bigger*	RP	particle	*up, off*	#	pound sign	#
JJS	superlative adj	*wildest*	SYM	symbol	+,%, &	"	left quote	' or "
LS	list item marker	*1, 2, One*	TO	"to"	*to*	"	right quote	' or "
MD	modal	*can, should*	UH	interjection	*ah, oops*	(	left paren	[, (, {, <
NN	sing or mass noun	*llama*	VB	verb base form	*eat*	)	right paren	],), }, >
NNS	noun, plural	*llamas*	VBD	verb past tense	*ate*	,	comma	,
NNP	proper noun, sing.	*IBM*	VBG	verb gerund	*eating*	.	sent-end punc	. ! ?
NNPS	proper noun, plu.	*Carolinas*	VBN	verb past part.	*eaten*	:	sent-mid punc	: ; ... – -

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

http://www.surdeanu.info/mihai/teaching/ista555-fall13/readings/PennTreebankTagset.html

Q1a

- Unigram (W_i)
- Rule-based
- HMM
- Classifier



Q1b

Pierre	Vinken	,	61	years	old	,	will	join	the	board	as	a	nonexecutive	director	Nov	29	.
		,				,											.

- NN — sing or mass noun — *llama*
- NNS — noun, plural — *llamas*
- NNP — proper noun, sing. — IBM
- VB — verb base form — *eat*
- JJ — adjective — *yellow*
- MD — modal — *can, should*
- CD — cardinal number — *one, two*
- DT — determiner — *a, the*
- IN — preposition/ subordin-conj — *of, in, by*

Q1b

Pierre	**Vinken**	**,**	**61**	**years**	**old**	**,**	**will**	**join**	**the**	**board**	**as**	**a**	**nonexecutive**	**director**	**Nov**	**29**	**.**
NNP	NNP	,	CD	NNS	JJ	,	MD	VB	DT	NN	IN	DT	JJ	NN	NNP	CD	.

Q2

2. Name the key differences and similarities between n-gram language models versus feed-forward neural language models.

Q2

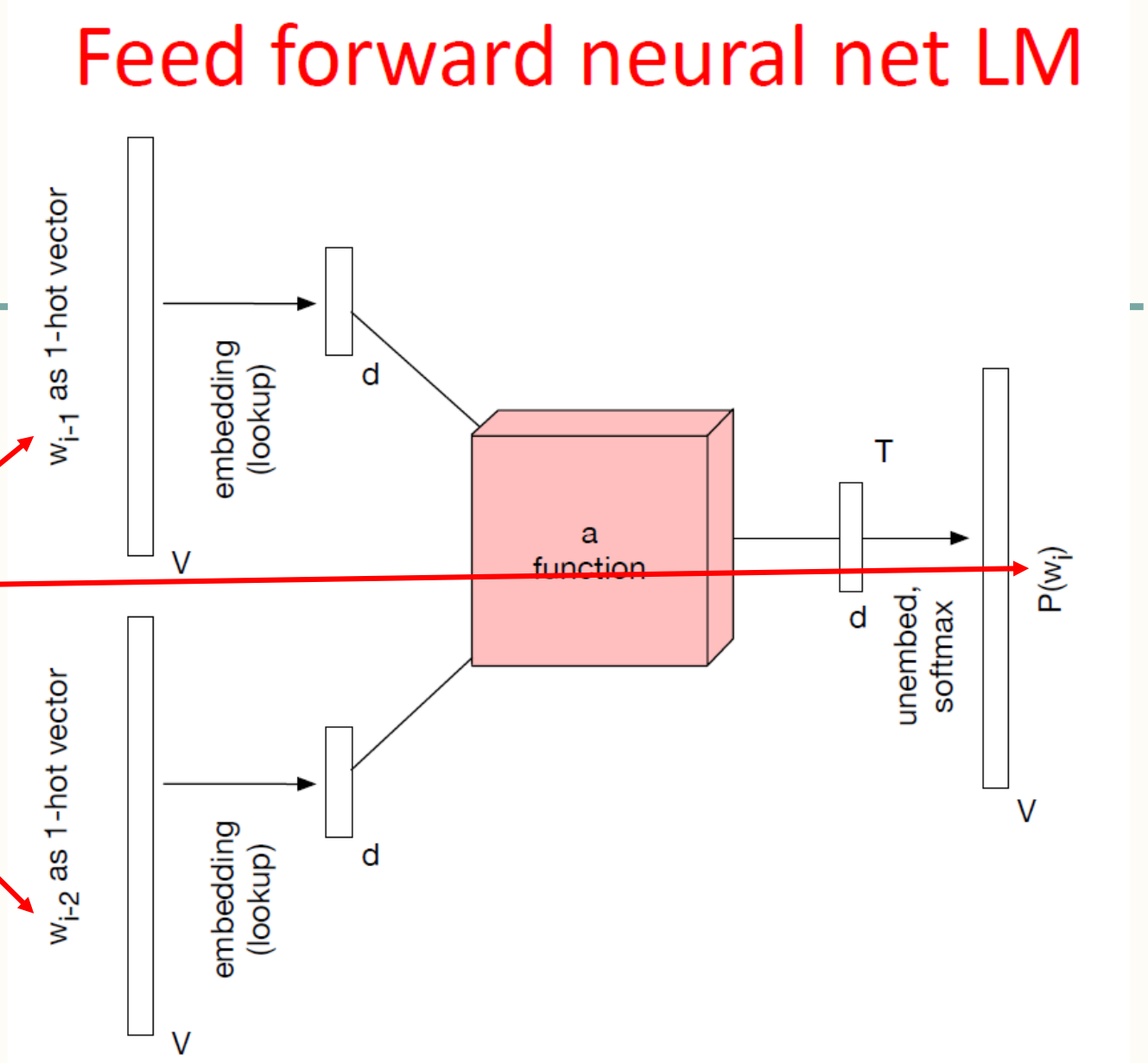
– **Language Model**

E.g., n = 3, a trigram model

$$P(w_1, w_2, .. w_m) = \prod_{i=1}^{m} P(w_i | w_{i-2} w_{i-1})$$

Markov Chain

Feed forward neural net LM

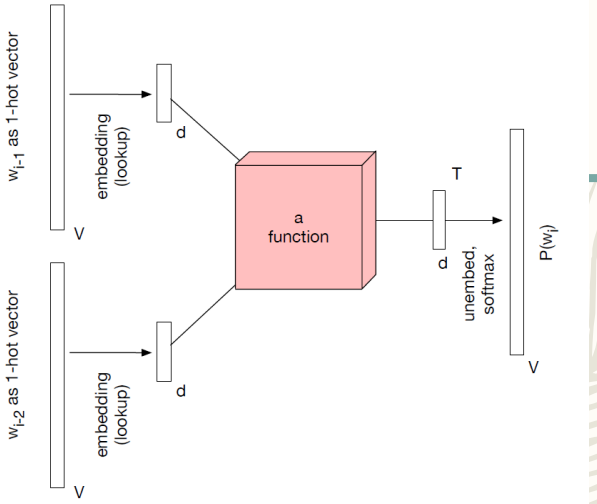


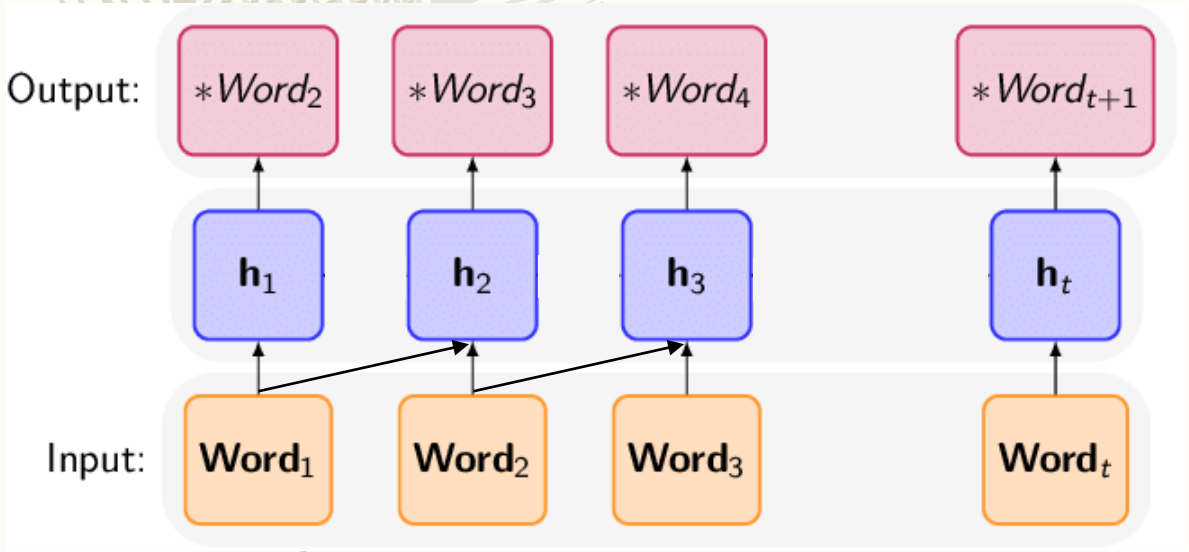
Q3

3. What does **recurrent** mean in the context of a recurrent neural network (RNN) language model? How does the approach differ from a feed-forward language model?

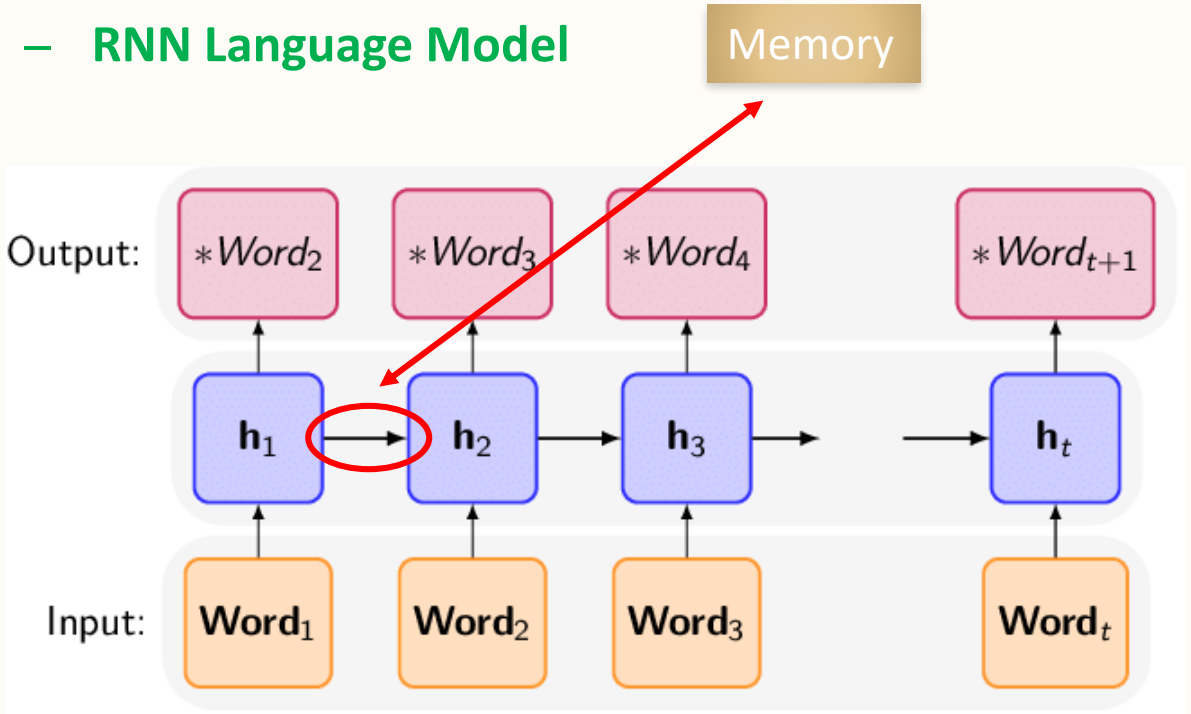
Q3

Feed forward neural net LM





– **RNN Language Model**



Q4

4. What advantage does a RNN language model have over a feed-forward language model?

Q4

- RNNLM can capture *long-distance dependencies*, while FFLM cannot.
- For example, it can balance quotes and brackets over long distances.
  - (..... (.... (..) ....) .....)